

(19) World Intellectual Property  
Organization  
International Bureau



(43) International Publication Date  
2 June 2005 (02.06.2005)

PCT

(10) International Publication Number  
**WO 2005/050473 A2**

(51) International Patent Classification<sup>7</sup>: **G06F 17/20**

[DE/DE]; c/o Philips Intellectual Property & Standards GmbH, Weissshausstr. 2, 52066 Aachen (DE).

(21) International Application Number:  
PCT/TB2004/052406

(74) Agents: MEYER, Michael et al.; Philips Intellectual Property & Standards GmbH, Weissshausstr. 2, 52066 Aachen (DE).

(22) International Filing Date:  
12 November 2004 (12.11.2004)

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
03104317.7 21 November 2003 (21.11.2003) EP

(71) Applicant (for DE only): PHILIPS INTELLECTUAL PROPERTY & STANDARDS GMBH [DE/DE]; Stein-damm 94, 20099 Hamburg (DE).

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

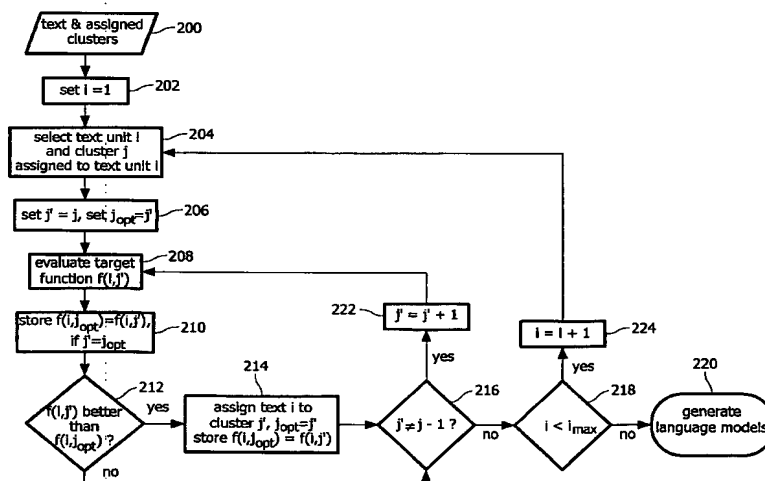
(71) Applicant (for all designated States except DE, US): KONINKLIJKE PHILIPS ELECTRONICS N. V. [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventor; and

(75) Inventor/Applicant (for US only): PETERS, Jochen

[Continued on next page]

(54) Title: CLUSTERING OF TEXT FOR STRUCTURING OF TEXT DOCUMENTS AND TRAINING OF LANGUAGE MODELS



(57) Abstract: The present invention relates to a method, a text segmentation system and a computer program product for clustering of text into text clusters representing a distinct semantic meaning. The text clustering method identifies text portions and assigns text portions to different clusters in such a way that each text cluster refers to one or several semantic topics. The clustering method incorporates an optimization procedure based on a re-clustering procedure evaluating a target function being indicative of the correlation between a text unit and a cluster. The text clustering method makes use of a text emission model and a cluster transition model and makes further use of various smoothing techniques.



**Published:**

— without international search report and to be republished  
upon receipt of that report

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*